Published online 19 November 2013 *Nucleic Acids Research, 2014, Vol. 42, Database issue* ***D7–D17***
doi:10.1093/nar/gkt1146

Database resources of the National Center for Biotechnology Information

NCBI Resource Coordinators*†

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received September 18, 2013; Revised October 23, 2013; Accepted October 24, 2013

ABSTRACT

In addition to maintaining the GenBank® nucleic acid sequence database, the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov) provides analysis and retrieval resources for the data in GenBank and other biological data made available through the NCBI Web site. NCBI resources include Entrez, the Entrez Programming Utilities, MyNCBI, PubMed, PubMed Central, PubReader, Gene, the NCBI Taxonomy Browser, BLAST, BLAST Link, Primer-BLAST, COBALT, RefSeq, UniGene, HomoloGene, ProtEST, dbMHC, dbSNP, dbVar, Epigenomics, the Genetic Testing Registry, Genome and related tools, the Map Viewer, Trace Archive, Sequence Read Archive, BioProject, BioSample, ClinVar, MedGen, HIV-1/Human Protein Interaction Database, Gene Expression Omnibus, Probe, Online Mendelian Inheritance in Animals, the Molecular Modeling Database, the Conserved Domain Database, the Conserved Domain Architecture Retrieval Tool, Biosystems, Protein Clusters and the PubChem suite of small molecule databases. Augmenting many of the Web applications are custom implementations of the BLAST program optimized to search specialized data sets. All these resources can be accessed through the NCBI home page.

INTRODUCTION

The National Center for Biotechnology Information (NCBI) at the National Institutes of Health (NIH) was created in 1988 to develop information systems for molecular biology. In addition to maintaining the GenBank® (1) nucleic acid sequence database, which receives data through the international collaboration with DDBJ and EMBL-Bank as well as from the scientific community, NCBI provides data retrieval systems and computational resources for the analysis of GenBank data and many other kinds of biological data. This article provides a summary of recent developments, including both new and updated resources, followed by an introduction to the Entrez system and a brief review of the suite of NCBI resources. All resources discussed are available from the NCBI Guide at www.ncbi.nlm.nih.gov and can also be located using the 'NCBI Web Site' database available in Entrez search menus. In most cases, the data underlying these resources and executables for the software described are available for download at ftp.ncbi.nlm.nih.gov.

RECENT DEVELOPMENTS

ClinVar

ClinVar (www.ncbi.nlm.nih.gov/clinvar/) is a new medical genetics resource that collects assertions of the relationships between human sequence variations and phenotypes (2). Submissions to ClinVar may specify the variation, the phenotype, the interpretation of the medical importance of the variation, the date that interpretation was last evaluated and the evidence supporting that interpretation, along with information about the submitter. ClinVar then integrates data from multiple submitters, and adds value such as HGVS nomenclature (3), identifiers in dbSNP or dbVar, calculation of molecular consequence and standardized terms. ClinVar also maps the names of submitted phenotypes into concepts that are integrated into MedGen. In addition to the assertion itself, ClinVar also reports the level of confidence in the assertion based on the number and type of submissions. Each of the individual assertions submitted to ClinVar has a unique accession of the format SCV000000000.0, and submissions that relate the same variant and phenotype are collected in reference records with accessions RCV000000000.0. Currently, ClinVar contains more than 50 000 reference records involving more than 15 000 genes. For more information about ClinVar, we would refer readers to its more complete description (2).

*To whom correspondence should be addressed. Eric W. Sayers. Tel: +1 301 496 2475; Fax: +1 301 480 9241; Email: sayers@ncbi.nlm.nih.gov
†The members of the NCBI Resource Coordinators group are listed in the Appendix.

MedGen

MedGen (www.ncbi.nlm.nih.gov/medgen/) is a new portal that collects information about human disorders having a genetic component. Launched in 2012, MedGen is developing into an important node in the NCBI suite of databases by standardizing how NCBI represents human phenotypes and by supporting communication among the databases and tools that depend on those standards (e.g. GTR, ClinVar, dbGaP and PheGenI). MedGen organizes information about phenotypes around a stable identifier assigned to terms used to name disorders and their clinical features. Whenever possible, this identifier is the same used by the Unified Medical Language System (www.nlm.nih.gov/research/umls/). MedGen uses a combination of automatic processing and curation to aggregate these data, and presents the results as a text report with several sections. These sections may include, depending on the available data, descriptions of the disease and its clinical features along with collections of relevant professional guidelines, clinical studies and systematic reviews. The reports may also include metadata and identifiers from other databases or ontologies such as the Human Phenotype Ontology (www.human-phenotype-ontology.org/), OMIM® (www.omim.org) and SNOMED CT (www.ihtsdo.org/snomed-ct/). In addition, MedGen provides access to genetic tests in the NIH Genetic Testing Registry (GTR) and to relevant records in other NCBI resources including PubMed, Gene and ClinVar. MedGen supports queries by terms, database identifiers such as MIM or SNOMED CT identifiers and concept relationships such as disorders sharing a clinical feature. In addition to the web service, MedGen is available through the Entrez Programming Utilities (E-utilities) (see below) and FTP (ftp.ncbi.nlm.nih.gov/pub/medgen).

PubReader

NCBI now offers a new, reader-friendly display option for viewing full-text articles in the PubMed Central (PMC) database. This new format, called PubReader, leverages features of HTML5 and CSS3 to address usability challenges specific to reading research articles on tablets and other small-screen devices. PubReader also works well on laptops and desktop machines. PubReader can display any PMC article that is available in the full-text HTML format, and automatically breaks the text into multiple columns and pages depending on the user's screen size. When viewed on a touch screen device, finger swipes and taps allow easy pagination; on devices with keyboards, the arrow keys provide similar functions. An image strip appears at the bottom of the display and shows thumbnail images of any figures and tables in the paper. Clicking or tapping a thumbnail opens a larger view of the content, and the image strip remains in place so that such access is available from any page in the article. This allows the reader to view any figure or table without losing their place in the article. More information about PubReader is available at the PubReader About page (www.ncbi.nlm.nih.gov/pmc/about/pubreader/).

SciENcv

Science Experts Network Curriculum Vitae (SciENcv) is a new tool in the My NCBI suite (https://www.ncbi.nlm.nih.gov/account/). SciENcv is designed to benefit research scientists and other individuals who apply for, receive or are associated with research investments from the NIH. SciENcv helps users to construct and maintain a profile for the purpose of creating a biographical sketch to be used in competing and non-competing federal grant applications at the NIH. A primary goal of the system is to reduce the administrative burden for researchers and institutions during the grant application process by streamlining the task of creating a biographical sketch and leveraging data that may exist for the user in eRA Commons as the Personal Profile Summary and NIH awards. SciENcv incorporates data from My Bibliography, a service that NIH-funded scientists use to manage and report publication data and that already integrates with a number of other databases such as PubMed, PMC, the NIH Manuscript Submission System (NIHMS), SPIRES and eRA Commons. Users have the option to keep their profile private or make it publicly available and share it with colleagues.

PubMed updates

Responding to requests from the community, PubMed abstracts now include keywords submitted by authors as well as non-English abstracts, if the publishers supplied these data. The additional language views are available through links on the Abstract display, where bold text indicates the language currently displayed. The PubMed search interface also includes new enhancements. The search menu includes a new section at the top that lists the last four Entrez databases searched. In addition, the PubMed Limits page has been recast as a set of sidebar filters on the left side of the search results page, allowing easier access to these search tools. Such sidebar filters have also been added to other Entrez databases, including NLM Catalog and ClinVar, and more will be added in the coming months. A new collection named 'Favorites' is now available for all My NCBI users, and an 'Add to Favorites' button in the right column of the abstract view provides a simple way to add any PubMed record to this collection. The 'Send to' menu also has a new option, 'Citation manager' that downloads PubMed records in a format suitable for importing into popular citation management programs (e.g. EndNote, Reference Manager). Finally, on the Advanced Search page, users can now download their search history as a CSV file, making it easy to save a record of a complex series of searches.

BLAST updates

In 2012, NCBI redesigned the BLAST report pages, adding new download options and improving navigation and performance. The 'Descriptions' section of the report now includes a download menu that provides direct access to FASTA data, either of the complete subject sequence or only the aligned portion, as well as to other formats including GenBank flat files, hit tables and XML.

Checkboxes to the left of each subject sequence allow users to download any arbitrary subset of the matching sequences. Clicking the title of a subject now links to the alignment display, which also contains similar download functions for that alignment along with controls to navigate between alignments and back to the 'Descriptions' table. The 'Alignments' section now initially displays only a few alignments, with additional alignments loaded on demand, greatly reducing the loading times for most searches.

In November 2012, the nucleotide collection database (nt) became the default search database for nucleotide BLAST. This popular database consists of sequences from databases in the International Nucleotide Sequence Database Collaboration (INSDC) and RefSeq sequences but excludes expressed sequence tags (EST), sequence-tagged sites (STS), genome survey sequences (GSS), whole-genome shotgun sequences (WGS), transcriptome shotgun assembly sequences and patent sequences as well as phase 0, 1 and 2 high-throughput genomic sequences. The nt database also recently joined a group of databases that support fast, indexed searches. The other databases that support such searches include the human G + T (genome plus transcript) and mouse G + T databases, as well as the human and mouse reference genome databases. These indexed searches use an in-memory index (4) and are available when using the megaBLAST algorithm. Such indexed megaBLAST searches at NCBI can search a nucleotide query of ~2000 bases against the nt database (43 billion bases) in a few seconds.

Both the microbial genomes BLAST service and the Sequence Read Archive (SRA) BLAST service received major updates in the past year. Both services now have interfaces that are similar to those of the traditional BLAST services, while offering several enhancements of particular use for these applications. The microbial interface includes an organism limit with an auto-complete feature that allows users to include or exclude one or more taxa. In addition, an option labeled 'Representative genomes only' provides access to a smaller and less redundant database that is especially helpful for bacterial species represented by numerous strains. These representative genomes are chosen either by the community or by NCBI, and they are typically well-studied strains that have garnered particular interest in the research community for that species. The SRA BLAST interface also provides an auto-complete input that allows users to specify an SRX accession, an SRX title or an organism name as the target database. Users can now search one or more SRX data sets in a combined database of up to 2 million reads. The available datasets include reads from Roche 454 and newer Illumina instruments (HiSeq and MiSeq) given the longer read lengths from these technologies.

Gene updates

The Gene resource at NCBI has continued to grow in size (over 13 000 000 records), in scope (more than 11 000 taxa) and in features and reports. The Table of Contents section in the upper right of the Gene full report now has subheadings, making it easier to jump to a subsection of the record in one click. It also alerts users to the types of content that are available for that record, such as the listing of locus-specific databases for human genes. In coordination with the GTR, ClinVar and MedGen projects, the phenotype section has been enhanced to facilitate exploration of human gene–disease relationships. The representation of neighboring genes, long displayed in the 'Genomic Context' section of the full report, is now available on the Gene FTP site, through the E-utilities using the new Gene neighbors link (gene_gene_neighbors). A summary of these improvements is available from the GeneNews RSS feed (www.ncbi.nlm.nih.gov/feed/rss.cgi?ChanKey = genenews).

Genome updates

The Genome database home page (www.ncbi.nlm.nih.gov/genome/) now provides links to two new pages that collect information about NCBI genome annotation pipelines for prokaryotes and eukaryotes, respectively. The prokaryotic page (www.ncbi.nlm.nih.gov/genome/annotation_prok/) provides a summary of the pipeline for prokaryotes along with links to more detailed information and instructions on how to run the pipeline on genome data being submitted to GenBank. The eukaryotic page (www.ncbi.nlm.nih.gov/genome/annotation_euk/) similarly provides an overview of the annotation pipeline, but importantly provides links to a page that lists annotation runs that are in progress and those that are complete, along with links to the data on the NCBI FTP site.

RefSeq updates

In July 2013, the Reference Sequence (RefSeq) project celebrated the 10th anniversary of comprehensive FTP releases, ending a decade during which the total amount of data in the collection (in base pairs and residues) increased by over 60-fold to 305 billion nucleotide bases and 11 billion amino acid residues. As RefSeq continues to respond to changes in the nature of sequence data and the needs of the community, in 2012 the project released a new accession type for proteins, the non-redundant WP records. NCBI created these WP records (with accessions WP_nnnnnnnnn) in response to the anticipated growth in the number of highly similar prokaryotic genomes resulting from clinical samples, for example, from outbreaks associated with a food-borne illness. These new records represent a unique protein sequence that may be annotated on multiple genomes, either from different strains of the same species and/or from different species. Thus, RefSeq will no longer assign unique GI numbers or accessions to proteins annotated on these new genomes; rather these proteins will have links to the WP record for that unique sequence. More information about these sequences can be found in this NCBI News article: www.ncbi.nlm.nih.gov/news/06-11-2013-wp-refseqs/.

Taxonomy updates

The Taxonomy database is now including type material for prokaryotic type strains and eukaryotic type specimens. Most of these new designations are currently available for prokaryotes and can be found on the summary page of the species. For example, several type strains are indicated for *Escherichia coli* (www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=562&lvl=0). In the near future, additional indexes will be added to Entrez so that users can easily retrieve all sequences derived from type strains for a species.

THE NCBI GUIDE AND THE ENTREZ SYSTEM

The NCBI Guide

The NCBI Guide serves not only as the NCBI home page but also as an interactive directory of the NCBI site. On the main page of the NCBI Guide, the categories in the Resource menu in the standard header are duplicated in a list on the left side of the page. Clicking on any category displays a list of relevant resources sorted into four groups: databases, downloads, submissions and tools. A list of how-to guides is also available via the 'How-To' tab on these pages. Popular resources are listed on the right under a 'Quick Links' heading, and on the main Guide page, a list of the most frequently used resources is provided in the 'Popular Resources' box and also as a list in the standard footer.

Entrez databases

Entrez (5) is an integrated database retrieval system that provides access to a diverse set of 42 databases that together contain 990 million records (Table 1). Links to the web portal for each of these databases are provided on the Entrez GQuery page (www.ncbi.nlm.nih.gov/gquery/). Entrez supports text searching using simple Boolean queries, downloading of data in various formats and linking of records between databases based on asserted relationships. In their simplest form, these links may be cross-references between a sequence and the abstract of the paper in which it is reported or between a protein sequence and its coding DNA sequence or its three-dimensional (3D) structure. Computationally derived links between 'neighboring records', such as those based on computed similarities among sequences or among PubMed abstracts, allow rapid access to groups of related records. Several popular links are displayed as Discovery Components in the right column of Entrez search result or record view pages, making these connections easier to find and explore. The LinkOut service expands the range of links to include external resources such as organism-specific genome databases. The records retrieved in Entrez can be displayed in many formats and downloaded singly or in batches.

Data sources and collaborations

NCBI receives data from three sources: direct submissions from external investigators, national and international collaborations or agreements with data providers and research consortia and internal curation efforts. The 'Data Source' column in Table 1 indicates those mechanisms by which each Entrez database receives data. More information about the various collaborations, agreements and curation efforts are available through the home pages of the individual resources.

Entrez Programming Utilities

The E-utilities constitute the Application Programming Interface (API) for the Entrez system. The API includes nine programs that support a uniform set of parameters used to search, link and download data from the Entrez databases. EInfo provides basic statistics on a given database, including the last update date and lists of all search fields and available links. ESearch returns the identifiers of records that match an Entrez text query, and when combined with EFetch or ESummary, provides a mechanism for downloading the corresponding data records. ELink gives users access to the vast array of links within Entrez so that data related to an input set can be retrieved. By assembling URL or SOAP calls to the E-utilities within simple scripts, users can create powerful applications to automate Entrez functions to accomplish batch tasks that are impractical using web browsers. The newest member of the E-utilities suite is ECitMatch, an API to the PubMed Batch Citation Matcher that returns PubMed IDs associated with citation strings. Other recent updates include version 2.0 of EInfo and support for downloading BioProject data using EFetch. Detailed documentation for using the E-utilities is available at eutils.ncbi.nlm.nih.gov.

LITERATURE

PubMed

The PubMed database contains citations from life science journals, many of which include abstracts and links to their full text articles.

PubMed Central

PMC (6) contains the full text of peer-reviewed journal articles in the life sciences and is the repository for all manuscripts arising from research using NIH funds and submitted through the NIHMS. Publishers that participate in PMC are required to provide free access to this full text either immediately after publication or within a 12-month period.

NLM Catalog

The NLM Catalog contains bibliographic data for the various items in the NLM collections, including journals, books, audiovisuals, computer software, electronic resources and other materials.

Medical Subject Headings

The Medical Subject Headings (MeSH) database (7) includes information about the NLM controlled vocabulary thesaurus used for indexing PubMed citations and

Table 1. The Entrez databases (as of 3 September 2013)

Database	Records	Section within this article	Data source
NCBI Web Site	21 929	Introduction	N
PubMed	23 052 796	Literature	C
PMC	2 836 592	Literature	D, C
NLM Catalog	1 485 089	Literature	C, N
MeSH	243 770	Literature	N
Books	222 232	Literature	C, N
Taxonomy[a]	1 153 795	Taxonomy	C, N
Nucleotide[a]	101 599 766	DNA and RNA	D (GenBank), C, N
EST[a]	74 911 096	DNA and RNA	D (GenBank)
GSS[a]	36 959 049	DNA and RNA	D (GenBank)
BioSample	2 100 817	DNA and RNA	N
SRA[a]	475 684	DNA and RNA	D
PopSet[a]	183 110	DNA and RNA	D (GenBank)
Protein[a]	94 102 424	Proteins	C, N
Protein Clusters[a]	382 691	Proteins	N
GEO Profiles[a]	91 392 791	Genes and expression	D
Probe	31 367 498	Genes and expression	D
Gene[a]	14 167 800	Genes and expression	C, N
UniGene[a]	6 467 085	Genes and expression	N
GEO Datasets[a]	1 044 344	Genes and expression	N
Biosystems[a]	522 277	Genes and expression	C
Homologene[a]	133 548	Genes and expression	N
Clone[a]	33 135 797	Genomes	D, N
UniSTS[a]	545 913	Genomes	D (dbSTS)
BioProject[a]	98 358	Genomes	D
Assembly	17 707	Genomes	C, N
Genome[a]	10 929	Genomes	C, N
MedGenEpigenomics[a]	10 811	Genomes	D
SNP[a]	300 258 943	Genetics and medicine	D (dbSNP), N
dbVar[a]	3 584 019	Genetics and medicine	D
MedGen[a]	169 433	Genetics and medicine	C, N
dbGaP	154 971	Genetics and medicine	D
ClinVar[a]	49 040	Genetics and medicine	D, N
PubMed Health	41 262	Genetics and medicine	C
GTR[a]	29 212	Genetics and medicine	D
OMIA	2844	Genetics and medicine	C
PubChem Substance[a]	119 813 846	Chemicals and bioassays	D
PubChem Compound[a]	47 757 896	Chemicals and bioassays	N
PubChem Bioassay[a]	717 429	Chemicals and bioassays	D
Structure[a]	92 993	Domains and structures	C, N
CDD[a]	48 034	Domains and structures	C, N

[a]Indicates that the data in this resource are available by FTP.
D, direct submission; C, collaboration/agreement; N, internal NCBI/NLM curation.

provides an interface for constructing PubMed queries using MeSH terms.

NCBI Bookshelf

The NCBI Bookshelf is an online service of the National Library of Medicine Literature Archive (NLM LitArch) that provides free access to the full text of books, reports, databases and documentation in the life sciences and healthcare fields.

TAXONOMY

The NCBI taxonomy database is a central organizing principle for the Entrez biological databases and provides links to all data for each taxonomic node, from superkingdoms to subspecies (8). The taxonomy database reflects sequence data from virtually all the formally described species of prokaryotes, and ~10% of the eukaryotes. The Taxonomy Browser (www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi) can be used to view the taxonomy tree or retrieve data from any of the Entrez databases for a particular organism or group.

DNA AND RNA

RefSeq

The RefSeq database (9) is a non-redundant set of curated and computationally derived sequences for transcripts, proteins and genomic regions. RefSeq DNA and RNA sequences can be searched and retrieved from the Nucleotide database, and the complete RefSeq collection is available in the RefSeq directory on the NCBI FTP site.

GenBank and other sources

GenBank (1) is the primary nucleotide sequence archive at NCBI and is a member of the INSDC. Sequences from

GenBank are available from three Entrez databases: Nucleotide, EST and GSS (specified as nuccore, nucest and nucgss within the E-utilities). The Nucleotide database contains all GenBank sequences except those within the EST or GSS GenBank divisions. The database also contains WGS sequences, Third Party Annotation (TPA) sequences and sequences imported from the Structure database.

PopSet

The PopSet database is a collection of related sequences and alignments derived from population, phylogenetic, mutation and ecosystem studies that have been submitted to GenBank. When available, PopSet alignments are shown in an embedded viewer on the PopSet record page.

Sequence Read Archive

SRA (10) is a repository for raw sequence reads and alignments generated by the latest generation of high-throughput nucleic acid sequencers. Data are deposited into SRA as supporting evidence for a wide range of study types including, *de novo* genome assemblies, genome wide association studies (GWAS), single nucleotide polymorphism and structural variation analysis, pathogen identification, transcript assembly, metagenomic community profiling and epigenetics.

Trace Archive

The Trace Archive contains sequence traces from gel and capillary electrophoresis sequencers. These data arise from whole genomes of pathogens, organismal shotgun and BAC clone projects and EST libraries. The Trace Assembly Archive is a companion resource that contains placements of individual trace reads on a GenBank sequence.

BioSample

The BioSample database provides annotation for biological samples used in a variety of studies submitted to NCBI, including genomic sequencing, microarrays, GWAS and epigenomics (11). The primary aim of BioSample is to address inconsistent annotations between similar samples from different studies so that investigators can more easily make connections between all the available data for a particular sample.

PROTEINS

RefSeq

In addition to genomic and transcript sequences, the RefSeq database (9) contains protein sequences that are curated and computationally derived from these DNA and RNA sequences. RefSeq protein sequences can be searched and retrieved from the Protein database, and the complete RefSeq collection is available in the RefSeq directory on the NCBI FTP site.

GenBank and other sources

As part of standard submission procedures, NCBI produces conceptual translations for any sequence in GenBank that contains a coding sequence and places these protein sequences in the Protein database. In addition to these 'GenPept' sequences, the Protein database also contains sequences from TPA, UniProtKB/Swiss-Prot (12), the Protein Research Foundation (PRF) and the Protein Data Bank (PDB) (13).

Protein Clusters

The Protein Clusters database contains sets of almost identical RefSeq proteins encoded by complete genomes from prokaryotes, eukaryotic organelles (mitochondria and chloroplasts), viruses and plasmids as well as from some protozoans and plants. The clusters are organized in a taxonomic hierarchy and are created based on reciprocal best-hit protein BLAST scores (14).

HIV-1/Human Protein Interaction Database

The HIV-1/Human Protein Interaction Database is an online presentation of documented interactions between HIV-1 proteins, host cell proteins, other HIV-1 proteins or proteins from disease organisms associated with HIV or AIDS (15). The Division of Acquired Immunodeficiency Syndrome of the National Institute of Allergy and Infectious Diseases in collaboration with the Southern Research Institute and NCBI, maintain these data.

BLAST SEQUENCE ANALYSIS

BLAST software

The BLAST programs (16–18) perform sequence-similarity searches against a variety of nucleotide and protein databases, returning a set of gapped alignments with links to full sequence records and related NCBI resources. The basic BLAST programs are also available as standalone command line programs, as network clients and as a local Web-server package at ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/ (Table 2).

BLAST databases

The default database for nucleotide BLAST searches (nr/nt) contains all RefSeq RNA records plus all GenBank sequences except for those from the EST, GSS, STS and HTG divisions. Another featured database is 'human genomic plus transcript' that contains human RefSeq transcript and genomic sequences arising from the NCBI annotation of the human genome. A similar database is available for mouse. Additional databases are also available and are described in links from the BLAST input form. Each of these databases can be limited to an arbitrary taxonomic node or those records satisfying any Entrez query.

For proteins the default database (nr) is a non-redundant set of all CDS translations from GenBank along with all RefSeq, UniProtKB/Swiss-Prot, PDB and PRF proteins. Subsets of this database are also available, such as the PDB or UniProtKB/Swiss-Prot sequences,

Table 2. Selected NCBI software available for download

Software	Available binaries	Category within this article
BLAST (standalone)	Win, Mac, LINUX, Solaris	BLAST sequence analysis
BLAST (network client)	Win, Mac, LINUX, Solaris	BLAST sequence analysis
BLAST (web server)	Mac, LINUX, Solaris	BLAST sequence analysis
CD-Tree	Win, Mac	Domains and structures
Cn3D	Win, Mac	Domains and structures
PC3D	Win, Mac, LINUX	Chemicals and bioassays
gene2xml	Win, Mac, LINUX, Solaris	Genes and expression
Genome Workbench	Win, Mac, LINUX	Genomes
splign	LINUX, Solaris	Genomes
tbl2asn	Win, Mac, LINUX, Solaris	Genomes

along with separate databases for sequences from patents and environmental samples. Like the nucleotide databases, these collections can be limited by taxonomy or an arbitrary Entrez query.

BLAST output formats

Standard BLAST output formats include the default pairwise alignment, several query-anchored multiple sequence alignment formats, an easily-parsable Hit Table and a report that organizes the BLAST hits by taxonomy. A 'pairwise with identities' mode better highlights differences between the query and a target sequence. A Tree View option for the Web BLAST service creates a dendrogram that clusters sequences according to their distances from the query sequence. Each alignment returned by BLAST is scored and assigned a measure of statistical significance, called the Expectation Value (E-value). The alignments returned can be limited by an E-value threshold or range.

Genomic BLAST

NCBI maintains Genomic BLAST services that mirror the design of the standard BLAST forms and allow users access to specialized databases for each particular genome. The default database contains the genomic sequence of an organism, but additional databases are provided depending on the available data and annotation. The default algorithm for Genomic BLAST is MegaBLAST (19), a faster version of standard nucleotide BLAST designed to find alignments between nearly identical sequences, typically from the same species. For rapid cross-species nucleotide queries, NCBI offers Discontiguous MegaBLAST, which uses a non-contiguous word match (20) as the nucleus for its alignments. Discontiguous MegaBLAST is far more rapid than a translated search such as blastx, yet maintains a competitive degree of sensitivity when comparing coding regions.

Primer-BLAST

Primer-BLAST is a tool for designing and analyzing PCR primers based on the existing program Primer3 (21) that designs PCR primers given a template DNA sequence. Primer-BLAST extends this functionality by running a BLAST search against a chosen database with the designed primers as queries, and then returns only those primer pairs specific to the desired target. If a user provides only one primer with the DNA template, the other primer will be designed and analyzed. If a user provides both primers and a template, the tool performs only the final BLAST analysis. If a user provides both primers but no template, primer-BLAST will display those templates that best match the primer pair. The available databases range from RefSeq mRNA or genomic sets for 1 of 12 model organisms to the entire BLAST nr database.

COBALT

COBALT (22) is a multiple alignment algorithm for proteins that finds a collection of pair-wise constraints derived from both the NCBI Conserved Domain Database (CDD) and the sequence similarity programs RPS-BLAST, BLASTp and PHI-BLAST. These pairwise constraints are then incorporated into a progressive multiple alignment. Links at the top of the COBALT report provide access to a phylogenetic tree view of the multiple alignment and allow users either to launch a modified search or download the alignment in several popular formats.

GENES AND EXPRESSION

Gene

Gene (23) provides an interface to curated sequences and descriptive information about genes with links to a wide variety of gene-related resources. These data are accumulated and maintained through several international collaborations in addition to curation by in-house staff. The complete Gene data set, as well as organism-specific subsets, is available in the compact NCBI Abstract Syntax Notation One (ASN.1) format on the NCBI FTP site. The gene2xml tool converts the native Gene ASN.1 format into XML and is available at ftp.ncbi.nlm.nih.gov/toolbox/ncbi_tools/converters/by_program/gene2xml/.

RefSeqGene

As part of the Locus Reference Genomic collaboration (www.lrg-sequence.org), RefSeqGene provides stable, standard human genomic sequences annotated with standard mRNAs for well-characterized human genes (9). RefSeqGene records are part of the RefSeq collection and are used to establish numbering systems for exons and introns and for reporting and identifying genomic variants, especially those of clinical importance (24). RefSeqGene records can be retrieved from the

Nucleotide database using the query 'refseqgene[keyword]', are available on corresponding Gene reports and can be downloaded from ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/RefSeqGene.

The Conserved CDS Database

The Conserved CDS Database (CCDS) project is a collaborative effort between NCBI, the European Bioinformatics Institute, the Wellcome Trust Sanger Institute (WTSI) and University of California, Santa Cruz to identify a set of human and mouse protein coding regions that are consistently annotated and of high quality (9). The collaborators prepare the CCDS set by comparing the annotations they have independently determined and then identifying those coding regions that have identical coordinates on the genome. Those regions that pass quality evaluations are then added to the CCDS set. The CCDS sequence data are available at ftp.ncbi.nlm.nih.gov/pub/CCDS/.

Gene Expression Omnibus

Gene Expression Omnibus (GEO) (25) is a data repository and retrieval system for high-throughput functional genomic data generated by microarray and next-generation sequencing technologies. In addition to gene expression data, GEO accepts data from studies of genome copy number variation, genome–protein interaction surveys and methylation profiling studies. The repository can capture fully annotated raw and processed data, enabling compliance with reporting standards such as 'Minimum Information About a Microarray Experiment' (23,24). GEO data are housed in two Entrez databases: GEO Profiles, which contains quantitative gene expression measurements for one gene across an experiment, and GEO DataSets, which contains entire experiments.

UniGene

UniGene (26) is a system for partitioning transcript sequences (including ESTs) from GenBank into a non-redundant set of clusters, each of which contains sequences that seem to be produced by the same transcription locus. UniGene clusters are created for all organisms for which there are 70 000 or more ESTs in GenBank.

HomoloGene

HomoloGene is a system that automatically detects homologs, including paralogs and orthologs, among the genes of 21 completely sequenced eukaryotic genomes. HomoloGene reports include homology and phenotype information drawn from Online Mendelian Inheritance in Man (27), Mouse Genome Informatics (28), Zebrafish Information Network (29), *Saccharomyces* Genome Database (30) and FlyBase (31). Information about the HomoloGene build procedure is provided at www.ncbi.nlm.nih.gov/HomoloGene/HTML/homologene_buildproc.html.

Probe

The Probe database is a registry of nucleic acid reagents designed for use in a wide variety of biomedical research applications including genotyping, SNP discovery, gene expression, gene silencing and gene mapping. Probe also includes information on reagent distributors, probe effectiveness and computed sequence similarities.

Biosystems

The Biosystems database collects together molecules represented in Gene, Protein and PubChem that interact in a biological system such as a biochemical pathway or disease. Currently, Biosystems receives data from the Kyoto Encyclopedia of Genes and Genomes (32–34), BioCyc (35), Reactome (36), the Pathway Interaction Database (37), WikiPathways (38,39) and Gene Ontology (40).

GENOMES

BioProject

The BioProject database is a central access point for metadata about research projects whose data are deposited in databases maintained by members of the INSDC. BioProject provides links to the primary data from these projects, which range from focused genome sequencing projects to large international collaborations with multiple sub-projects incorporating experiments resulting in nucleotide sequence sets, genotype/phenotype data, sequence variants or epigenetic information.

Genome Reference Consortium

The Genome Reference Consortium (GRC) (www.genomereference.org) is an international collaboration between the WTSI, the Genome Institute at Washington University, EMBL and NCBI that aims to produce assemblies of higher eukaryotic genomes that best reflect complex allelic diversity consistent with currently available data. The GRC currently produces assemblies for human (GRCh37), mouse (GRCm38) and zebrafish (Zv9). Between major assembly releases the GRC provides minor 'patch' releases that provide additional sequence scaffolds that either correct errors in the assembly (fix patches) or add an alternate loci (novel patches). GRC staff then incorporate these changes into the next major assembly release. GRC data are available for download from the NCBI FTP site (ftp.ncbi.nlm.nih.gov/pub/grc/), and assembly data are available from the GenBank genomes ftp site (ftp.ncbi.nlm.nih.gov/genbank/genomes/).

Clone Database

Clone Database (CloneDB) is a resource for finding descriptions, sources, map positions and distributor information about available clones and libraries (41). For both genomic and cell-based clones and libraries, CloneDB contains information about the sequences themselves, such as their genomic mapping positions and associated markers, along with details about how the libraries were constructed.

Epigenomics

The Epigenomics database collects data from studies examining epigenetic features such as post-translational

modifications of histone proteins, genomic DNA methylation, chromatin organization and the expression of non-coding regulatory RNA (42). The Epigenomics database provides displays ('genome tracks') of the raw data (stored in the GEO and SRA databases) mapped to genomic coordinates. Data from the Roadmap Epigenomics project, currently stored in GEO (www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/), are being mirrored and are available for viewing and downloading.

Influenza Genome Resources

The Influenza Virus Resource links genome sequence data from the Influenza Genome Sequencing Project (43) to the most recent scientific literature in PubMed on influenza as well as to population studies and protein sequences and structures. The NCBI Virus Variation resource extends these services to the dengue and West Nile viruses.

GENETICS AND MEDICINE

dbGaP

The Database of Genotypes and Phenotypes (dbGaP) (44) archives, distributes and supports submission of data that correlate genomic characteristics with observable traits. This database is a designated NIH repository for NIH-funded GWAS results (grants.nih.gov/grants/gwas/). To protect the confidentiality of study subjects, dbGaP accepts only de-identified data and requires investigators to go through an authorization process to access individual-level data. Study documents, protocols and subject questionnaires are available without restriction.

dbVar

The Database of Genomic Structural Variation (dbVar) is an archive of large-scale genomic variants (generally >50 bp) such as insertions, deletions, translocations and inversions (45). These data are derived from several methods including computational sequence analysis and microarray experiments.

dbSNP

The Database of Short Genetic Variations (dbSNP) (46) is a repository of all types of short genetic variations <50 bp in length, and so is a complement to dbVar. dbSNP accepts submissions of common as well as polymorphic variations, and contains both germline and somatic variations. In addition to archiving molecular details for each submission and calculating submitted variant locations on each genome assembly, dbSNP maintains information about population-specific allele frequencies and genotypes, reports the validation state of each variant and indicates if a variation call may be suspect because of paralogy (47).

OMIA

Online Mendelian Inheritance in Animals (OMIA) is a database of genes, inherited disorders and traits in animal species other than human and mouse and is authored by Professor Frank Nicholas of the University of Sydney, Australia and colleagues (48).

dbMHC, dbLRC and dbRBC

NCBI maintains three databases for routine clinical applications: dbMHC, dbLRC and dbRBC. dbMHC focuses on the Major Histocompatibility Complex (MHC) and contains sequences and frequency distributions for MHC alleles. dbMHC also contains HLA genotype and clinical outcome information on hematopoietic cell transplants performed worldwide. dbLRC offers a comprehensive collection of alleles of the leukocyte receptor complex with an emphasis on KIR genes. dbRBC provides data on genes for red blood cell antigens along with access to the International Society of Blood Transfusion allele nomenclature of blood group alleles. dbRBC also hosts the Blood Group Antigen Gene Mutation Database (49) and integrates it with resources at NCBI. All three databases dbMHC, dbLRC and dbRBC provide multiple sequence alignments, analysis tools to interpret homozygous or heterozygous sequencing results (50) and tools for DNA probe alignments.

CHEMICALS AND BIOASSAYS

PubChem

PubChem (51,52) is the informatics backbone for the NIH Roadmap Initiative on molecular libraries and focuses on the chemical, structural and biological properties of small molecules, in particular their roles as diagnostic and therapeutic agents. A suite of three Entrez databases, PCSubstance, PCCompound and PCBioAssay, contain the structural and bioactivity data of the PubChem project. PubChem also provides a diverse set of 3D conformers for 90% of the records in the PubChem Compound database.

DOMAINS AND STRUCTURES

Molecular Modeling Database

Molecular Modeling Database (MMDB) (53) contains experimentally determined coordinate sets from PDB (13) augmented with domain annotations and links to relevant literature, protein and nucleotide sequences, chemicals (PDB heterogens) and conserved domains in the CDD (54). MMDB also provides interactive views of the data in Cn3D (55), the NCBI structure and alignment viewer. MMDB provides structural neighbors for each record based on similarities computed by the VAST algorithm between compact structural domains within protein structures (56,57).

Conserved Domain Database

CDD (58) contains PSI-BLAST-derived Position Specific Score Matrices representing domains taken from the Simple Modular Architecture Research Tool (Smart) (59), Pfam (60), TIGRFAM (61), and from domain alignments derived from Clusters of Orthologous Groups (COGs) and Protein Clusters. In addition, CDD includes

superfamily records that contain sets of CDs from one or more source databases that generate overlapping annotation on the same protein sequences.

FOR FURTHER INFORMATION

The resources described here include documentation, other explanatory material and references to collaborators and data sources on their respective web sites. An alphabetical list of NCBI resources is available from a link above the category list on the left side of the NCBI home page. The NCBI Help Manual and the NCBI Handbook, both available as links in the common page footer, describe the principal NCBI resources in detail. The NCBI Education page (www.ncbi.nlm.nih.gov/Education/) lists links to documentation, tutorials and educational tools along with links to outreach initiatives including Discovery Workshops, webinars and upcoming conference exhibits. The Education page, along with the standard NCBI page footer, contains links to the NCBI YouTube channel that contains a variety of video tutorials. A user-support staff is available to answer questions at info@ncbi.nlm.nih.gov. Updates on NCBI resources and database enhancements are described on the NCBI News site (www.ncbi.nlm.nih.gov/news/), which also links to the NCBI social media sites (FaceBook and Twitter), the 'NCBI Insights' blog, and the several mailing lists and RSS feeds that provide updates on services and databases.

FUNDING

Funding for open access charge: Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Cavanaugh,M., Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2014) GenBank. *Nucleic Acids Res.*, **42**, D32–D37.
2. Landrum,M., Lee,J.M., Riley,G., Jang,W., Rubinstein,W.S., Church,D.M. and Maglott,D. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.
3. Antonarakis,S.E. (1998) Recommendations for a nomenclature system for human gene mutations. Nomenclature Working Group. *Hum. Mutat.*, **11**, 1–3.
4. Morgulis,A., Coulouris,G., Raytselis,Y., Madden,T.L., Agarwala,R. and Schaffer,A.A. (2008) Database indexing for production MegaBLAST searches. *Bioinformatics*, **24**, 1757–1764.
5. Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.
6. Sequeira,E. (2003) PubMed Central—three years old and growing stronger. *ARL*, **228**, 5–9.
7. Sewell,W. (1964) Medical subject headings in Medlars. *Bull. Med. Libr. Assoc.*, **52**, 164–170.
8. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
9. Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
10. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
11. Barrett,T., Clark,K., Gevorgyan,R., Gorelenkov,V., Gribov,E., Karsch-Mizrachi,I., Kimelman,M., Pruitt,K.D., Resenchuk,S., Tatusova,T. *et al.* (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.*, **40**, D57–D63.
12. Magrane,M. and Consortium,U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)*, **2011**, bar009.
13. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
14. Klimke,W., Agarwala,R., Badretdin,A., Chetvernin,S., Ciufo,S., Fedorov,B., Kiryutin,B., O'Neill,K., Resch,W., Resenchuk,S. *et al.* (2009) The National Center for Biotechnology Information's protein clusters database. *Nucleic Acids Res.*, **37**, D216–D223.
15. Fu,W., Sanders-Beer,B.E., Katz,K.S., Maglott,D.R., Pruitt,K.D. and Ptak,R.G. (2009) Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Res.*, **37**, D417–D422.
16. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
17. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
18. Ye,J., McGinnis,S. and Madden,T.L. (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res.*, **34**, W6–W9.
19. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
20. Ma,B., Tromp,J. and Li,M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.
21. Rozen,S. and Skalestsky,H.J. (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz,S. and Misener,S. (eds), *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Humana Press, Totowa, NJ, pp. 365–386.
22. Papadopoulos,J.S. and Agarwala,R. (2007) COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics*, **23**, 1073–1079.
23. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.
24. Gulley,M.L., Braziel,R.M., Halling,K.C., Hsi,E.D., Kant,J.A., Nikiforova,M.N., Nowak,J.A., Ogino,S., Oliveira,A., Polesky,H.F. *et al.* (2007) Clinical laboratory reports in molecular pathology. *Arch. Pathol. Lab. Med.*, **131**, 852–863.
25. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**, D991–D995.
26. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
27. Amberger,J., Bocchini,C.A., Scott,A.F. and Hamosh,A. (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, **37**, D793–D796.
28. Eppig,J.T., Blake,J.A., Bult,C.J., Kadin,J.A. and Richardson,J.E. (2007) The mouse genome database (MGD): new features facilitating a model system. *Nucleic Acids Res.*, **35**, D630–D637.
29. Sprague,J., Bayraktaroglu,L., Clements,D., Conlin,T., Fashena,D., Frazer,K., Haendel,M., Howe,D.G., Mani,P., Ramachandran,S. *et al.* (2006) The Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Res.*, **34**, D581–D585.
30. Hong,E.L., Balakrishnan,R., Dong,Q., Christie,K.R., Park,J., Binkley,G., Costanzo,M.C., Dwight,S.S., Engel,S.R., Fisk,D.G. *et al.* (2008) Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res.*, **36**, D577–D581.

31. Crosby,M.A., Goodman,J.L., Strelets,V.B., Zhang,P. and Gelbart,W.M. (2007) FlyBase: genomes by the dozen. *Nucleic Acids Res.*, **35**, D486–D491.
32. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
33. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
34. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
35. Keseler,I.M., Bonavides-Martinez,C., Collado-Vides,J., Gama-Castro,S., Gunsalus,R.P., Johnson,D.A., Krummenacker,M., Nolan,L.M., Paley,S., Paulsen,I.T. *et al.* (2009) EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res.*, **37**, D464–D470.
36. Matthews,L., Gopinath,G., Gillespie,M., Caudy,M., Croft,D., de Bono,B., Garapati,P., Hemish,J., Hermjakob,H., Jassal,B. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, **37**, D619–D622.
37. Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. and Buetow,K.H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674–D679.
38. Kelder,T., Pico,A.R., Hanspers,K., van Iersel,M.P., Evelo,C. and Conklin,B.R. (2009) Mining biological pathways using WikiPathways web services. *PLoS One*, **4**, e6447.
39. Pico,A.R., Kelder,T., van Iersel,M.P., Hanspers,K., Conklin,B.R. and Evelo,C. (2008) WikiPathways: pathway editing for the people. *PLoS Biol.*, **6**, e184.
40. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
41. Schneider,V.A., Chen,H.C., Clausen,C., Meric,P.A., Zhou,Z., Bouk,N., Husain,N., Maglott,D.R. and Church,D.M. (2013) Clone DB: an integrated NCBI resource for clone-associated data. *Nucleic Acids Res.*, **41**, D1070–D1078.
42. Fingerman,I.M., McDaniel,L., Zhang,X., Ratzat,W., Hassan,T., Jiang,Z., Cohen,R.F. and Schuler,G.D. (2011) NCBI Epigenomics: a new public resource for exploring epigenomic data sets. *Nucleic Acids Res.*, **39**, D908–D912.
43. Ghedin,E., Sengamalay,N.A., Shumway,M., Zaborsky,J., Feldblyum,T., Subbu,V., Spiro,D.J., Sitz,J., Koo,H., Bolotov,P. *et al.* (2005) Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature*, **437**, 1162–1166.
44. Manolio,T.A., Rodriguez,L.L., Brooks,L., Abecasis,G., Ballinger,D., Daly,M., Donnelly,P., Faraone,S.V., Frazer,K., Gabriel,S. *et al.* (2007) New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat. Genet.*, **39**, 1045–1051.
45. Church,D.M., Lappalainen,I., Sneddon,T.P., Hinton,J., Maguire,M., Lopez,J., Garner,J., Paschall,J., Dicuccio,M., Yaschenko,E. *et al.* (2010) Public data archives for genomic structural variation. *Nat. Genet.*, **42**, 813–814.
46. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
47. Sudmant,P.H., Kitzman,J.O., Antonacci,F., Alkan,C., Malig,M., Tsalenko,A., Sampas,N., Bruhn,L., Shendure,J. and Eichler,E.E. (2010) Diversity of human copy number variation and multicopy genes. *Science*, **330**, 641–646.
48. Lenffer,J., Nicholas,F.W., Castle,K., Rao,A., Gregory,S., Poidinger,M., Mailman,M.D. and Ranganathan,S. (2006) OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. *Nucleic Acids Res.*, **34**, D599–D601.
49. Blumenfeld,O.O. and Patnaik,S.K. (2004) Allelic genes of blood group antigens: a source of human mutations and cSNPs documented in the Blood Group Antigen Gene Mutation Database. *Hum. Mutat.*, **23**, 8–16.
50. Helmberg,W., Dunivin,R. and Feolo,M. (2004) The sequencing-based typing tool of dbMHC: typing highly polymorphic gene sequences. *Nucleic Acids Res.*, **32**, W173–W175.
51. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J. and Bryant,S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.
52. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J., Zhou,Z., Han,L., Karapetyan,K., Dracheva,S., Shoemaker,B.A. *et al.* (2012) PubChem's BioAssay Database. *Nucleic Acids Res.*, **40**, D400–D412.
53. Madej,T., Addess,K.J., Fong,J.H., Geer,L.Y., Geer,R.C., Lanczycki,C.J., Liu,C., Lu,S., Marchler-Bauer,A., Panchenko,A.R. *et al.* (2012) MMDB: 3D structures and macromolecular interactions. *Nucleic Acids Res.*, **40**, D461–D464.
54. Marchler-Bauer,A., Lu,S., Anderson,J.B., Chitsaz,F., Derbyshire,M.K., DeWeese-Scott,C., Fong,J.H., Geer,L.Y., Geer,R.C., Gonzales,N.R. *et al.* (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.*, **39**, D225–D229.
55. Wang,Y., Geer,L.Y., Chappey,C., Kans,J.A. and Bryant,S.H. (2000) Cn3D: sequence and structure views for Entrez. *Trends Biochem. Sci.*, **25**, 300–302.
56. Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.
57. Madej,T., Gibrat,J.F. and Bryant,S.H. (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.
58. Marchler-Bauer,A., Anderson,J.B., Chitsaz,F., Derbyshire,M.K., DeWeese-Scott,C., Fong,J.H., Geer,L.Y., Geer,R.C., Gonzales,N.R., Gwadz,M. *et al.* (2009) CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res.*, **37**, D205–D210.
59. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.
60. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
61. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.

APPENDIX

NCBI Resource Coordinators: Abigail Acland, Richa Agarwala, Tanya Barrett, Jeff Beck, Dennis A. Benson, Colleen Bollin, Evan Bolton, Stephen H. Bryant, Kathi Canese, Deanna M. Church, Karen Clark, Michael DiCuccio, Ilya Dondoshansky, Scott Federhen, Michael Feolo, Lewis Y. Geer, Viatcheslav Gorelenkov, Marilu Hoeppner, Mark Johnson, Christopher Kelly, Viatcheslav Khotomlianski, Avi Kimchi, Michael Kimelman, Paul Kitts, Sergey Krasnov, Anatoliy Kuznetsov, David Landsman, David J. Lipman, Zhiyong Lu, Thomas L. Madden, Tom Madej, Donna R. Maglott, Aron Marchler-Bauer, Ilene Karsch-Mizrachi, Terence Murphy, James Ostell, Christopher O'Sullivan, Anna Panchenko, Lon Phan, Don Preussm Kim D. Pruitt, Wendy Rubinstein, Eric W. Sayers, Valerie Schneider, Gregory D. Schuler, Edwin Sequeira, Stephen T. Sherry, Martin Shumway, Karl Sirotkin, Karanjit Siyan, Douglas Slotta, Alexandra Soboleva, Vladimir Soussov, Grigory Starchenko, Tatiana A. Tatusova, Bart W. Trawick, Denis Vakatov, Yanli Wang, Minghong Ward, W. John Wilbur, Eugene Yaschenko and Kerry Zbicz